(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification[7]:   C12Q 1/68

(21) International Application Number: PCT/US00/25464

(22) International Filing Date:
                14 September 2000 (14.09.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
  60/154,480   17 September 1999 (17.09.1999)   US
  09/528,414   17 March 2000 (17.03.2000)   US

(71) Applicant *(for all designated States except US)*:
AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: HU, Jing-Shan [CN/US]; 1247 Lakeside Dr. #3034, Sunnyvale, CA 94086 (US). DURST, Mark [US/US]; 980 Estudillo

Avenue, San Leandro, CA 94577 (US). KHURGIN, Elina [US/US]; 22999 Voss Avenue, Cupertino, CA 95014 (US). BALBAN, David, J. [US/US]; 10224 Peninsula Avenue, Cupertino, CA 95014 (US).
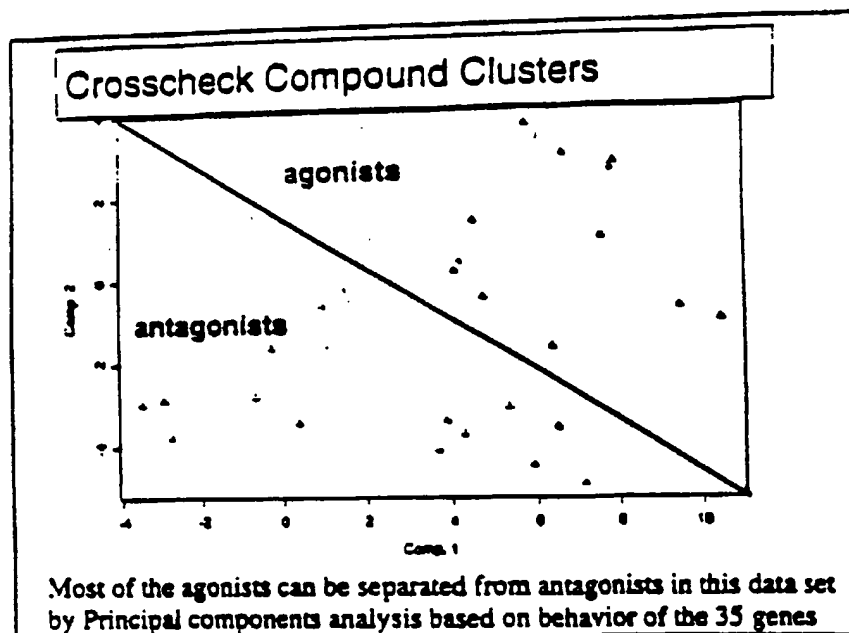
(74) Agents: LIEBESCHUETZ, Joe; Townsend and Townsend and Crew LLP, Two Embarcadero Center, 8th floor, San Francisco, CA 94111 et al. (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHOD OF CLUSTER ANALYSIS OF GENE EXPRESSION PROFILES

Crosscheck Compound Clusters

agonists

antagonists

Most of the agonists can be separated from antagonists in this data set by Principal components analysis based on behavior of the 35 genes

(57) Abstract: The present invention provides methods, systems and computer software products for gene expression data analysis. In one preferred embodiment, the dimension of gene expression profiles are reduced. The reduced gene expression profiles are then subjected to cluster analysis.

WO 01/20043 A1

METHOD OF CLUSTER ANALYSIS OF GENE EXPRESSION PROFILES

RELATED APPLICATION

This application claims the priority of U. S. Provisional Application, Serial

5   No. 60/154,480, attorney docket No. 3259, filed on September 17, 1999.   The 60/154,480

application is incorporated herein by reference for all purposes.

FIELD OF INVENTION

The present invention is related to biological data analysis methods and

10   computer program products.

BACKGROUND OF THE INVENTION

Many biological functions are carried out by regulating the expression levels

of various genes, either through changes in the copy number of the genetic DNA, through

15   changes in levels of transcription (*e.g.* through control of initiation, provision of RNA

precursors, RNA processing, *etc.*) of particular genes, or through changes in protein synthesis.

For example, control of the cell cycle and cell differentiation, as well as diseases, are

characterized by the variations in the transcription levels of a group of genes.

Recently, massive parallel gene expression monitoring methods have been

20   developed to monitor the expression of a large number of genes using nucleic acid array

technology which was described in detail in, for example, U.S. Patent Number 5,871,928; de

Saizieu, *et al.*, 1998, Bacteria Transcript Imaging by Hybridization of total RNA to

Oligonucleotide Arrays, NATURE BIOTECHNOLOGY, 16:45-48; Wodicka *et al.*, 1997,

Genome-wide Expression Monitoring in *Saccharomyces cerevisiae*, NATURE

25   BIOTECHNOLOGY 15:1359-1367; Lockhart *et al.*, 1996, Expression Monitoring by

Hybridization to High Density Oligonucleotide Arrays. NATURE BIOTECHNOLOGY 14:1675-

1680; Lander, 1999, Array of Hope, NATURE-GENETICS, 21(suppl.), at 3.

Massive parallel gene expression monitoring experiments generate

unprecedented amounts of information. A single hybridization experiment can produce

30   quantitative results for as many as 40,000 human genes. For example, a commercially

available GeneChip® array set is capable of monitoring the expression levels of

approximately 6,500 murine genes and expressed sequence tags (ESTs) (Affymetrix, Inc,

Santa Clara, CA, USA). It is desirable to use expression profiling to identify patterns or

fingerprints that signify specific drug effects, including side effects that can indicate toxicity. Therefore, there is a great need in the art for methods and computer program products to organize, access and analyze the vast amount of information collected using massive parallel gene expression monitoring methods.

5   Recently, cluster analysis has been applied to interpret gene expression data. Clustering of gene expression data has been shown to result in groups of genes that have related functions (*see, e.g.,* Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., Somogyi, R.: "Cluster analysis and data visualization of large-scale gene expression data", Pacific Symp. Biocomp. 98 3: 42-53, 1998; Eisen, M.B., Spellman, P.T., Brown, P.O.,

10   Botstein, D.: "Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. USA 95: 14863-14868, 1998; and To"ro"nen, P., Kolehmainen, M., Wong, G., Castre'n, E.: "Analysis of gene expression data using self-organizing maps", FEBS L. 451: 142-146, 1999). A number of computer software products have been released for performing cluster analysis of gene expression profiles (SpotFire Pro for Windows,

15   http://www.spotfire.com; DataDesk, http://www.datadesk.com; GeneCluster, www.genome.wi.mit.edu/MPR/software.html). However, before this invention, the cluster analysis methods and software products were not useful for analyzing gene expression profiles generated by non-toxic drug treatments. Cluster analysis performed on those profiles were often overwhelmed by noises. Therefore, there is a great need in the art for

20   methods, algorithms and computer program products for clustering genes based upon their expression during relatively mild environmental changes, such as under the treatment of non-toxic levels of drugs or drug candidate.

BRIEF DESCRIPTION OF THE DRAWINGS

25   Figure 1 illustrates an example of a computer system that may be utilized to execute the software of an embodiment of the invention.

Figure 2 illustrates a system block diagram of the computer system of Fig. 1.

30   Figure 3 illustrates one embodiment of a process for gene expression profile analysis.

Figure 4 illustrates one embodiment of an iterative process for gene expression profile analysis.

Figure 5 illustrates one embodiment of a process for reducing the dimension of gene expression profiles.

Figure 6 shows drug compounds and their properties.

Figure 7 shows a compound cluster.

Figure 8 shows a display of gene using principal components as axes.

SUMMARY OF THE INVENTION

The present invention provides methods and computer software (program) products for analyzing biological profiles. The methods and computer program products are particularly useful for analyzing gene expression profiles representing the state of a biological sample under various drug treatments. Methods and computer program products for efficiently displaying data analysis result is also provided.

In one aspect of the invention, methods and computer program products are provided to reduce and the dimension of biological profiles and to subject the reduced biological profiles to statistical analysis. In some embodiments of a method for analyzing a plurality of biological profiles, the number of biological variables, preferably more than 100, more preferably more than 1000 and most preferably more than 2000 variables, in the biological profiles are reduced to obtain a set of reduced profiles, and the reduced profiles are subjected to statistical analysis. In one particularly preferred embodiment, the dimension of the biological profiles are reduced by selecting variables based upon the degree of variation of the biological variables among the profiles. The measurement of the variation may be the variance of the biological variables among the profiles or the fold change of the biological variables among the profiles. In another preferred embodiment, the dimension of the biological profiles are reduced by selecting variables from the biological variables based upon the level of the biological variables among the profiles. In yet another embodiment, several methods described above may be combined to reduce the dimension of the profiles. In some

preferred embodiments, the statistical analysis is cluster analysis or principal component analysis. The cluster analysis may be a hierarchical cluster analysis.

In preferred embodiments, the biological profiles are gene expression profiles and each of the biological variables represents the expression of a gene. While this invention

5   is not limited to any particular method of measuring gene expression, gene expression is preferably measured using nucleic acid probe arrays.

In another aspect of the invention, computer program products are provided. A exemplary computer program product comprises: a) computer code that receives a plurality of biological profiles, each of the profiles comprises a plurality of biological

10   variables; b) computer code that reduces the number of the biological variables to obtain a set of reduced profiles; c) computer code that performs statistical analysis of the reduced profiles; and d) a computer readable medium that stores the computer codes. In some embodiments, the computer program product of the invention comprises computer code that selects the reduced set of variables from the biological variables based upon the degree of variation of

15   the biological variables among the profiles. The degree of variation may be the variance of the biological variables among the profiles or the fold change of the biological variables among the profiles. In some other embodiments, the computer program product of the invention contains computer code that selects the reduced set of variables from the biological variables based upon the level of the biological variables among said profiles. In one

20   particularly preferred embodiment, the computer program product of the invention contains code that selects variables using a combination of the methods. The computer program product of the invention may also contain code for cluster analysis (such as hierarchical cluster analysis) and /or for principal component analysis.

In an additional aspect of the invention, methods for studying a plurality of

25   drugs are provided. In some embodiments, the expression of more than 50 genes, preferably more than 1000 genes, more preferably more than 2000 genes and most preferably more than 4000 genes, in a biological sample in response to a plurality of drugs is measured, preferably by nucleic acid probe arrays, to obtain a plurality of gene expression profiles, each of the profiles representing the response of the biological sample to one of the plurality of drugs.

30   The dimension of the gene expression profiles is reduced by selecting a plurality of genes from the at least 50 genes to obtain a set of reduced gene expression profiles; and a statistical analysis is performed using said reduced gene expression profiles. Statistical analysis method may be a cluster analysis or a principal component analysis.

A hierarchical cluster analysis is preferred in some embodiments.

In one particularly preferred embodiment, a plurality of drugs are classified according to dimension reduction methods. A cluster analysis is performed using genes as variables (Y axis). Similarly, genes may be classified by a cluster analysis using drugs as variables (Y axis). Drugs or genes may also be classified according the dimension reduction method of the invention and a principal component analysis. In such an embodiment, the genes may be displayed in a surface comprising a first axis representing the first component of a principal analysis and a second axis representing the second component of the principal analysis. The first axis is perpendicular to the second axis.

In yet another aspect of the invention, computer program products are provided to display gene expression data in order to facilitate pattern identification. In some embodiments, a computer program product is provided. The computer program product comprises: a) computer code that receives a plurality of biological profiles, each of the profiles comprises a plurality of biological variables; b) computer code that reduces the number of the biological variables to obtain a set of reduced profiles; c) computer code that performs a principal component analysis of the reduced profiles; d) computer code that displays the biological variables according a first axis and a second axis, wherein the first axis is the first component of the principal component analysis and the second axis is the second component of the principal component analysis, and wherein the first axis is perpendicular to the second axis; and e) computer readable medium that stores the computer codes.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Reference will now be made in detail to the preferred embodiments of the invention. While the invention will be described in conjunction with the preferred embodiments, it will be understood that they are not intended to limit the invention to these embodiments. On the contrary, the invention is intended to cover alternatives, modifications and equivalents, which may be included within the spirit and scope of the invention. For example, the invention will be described by referring to embodiments providing methods, algorithms, data analysis systems and computer program products for discovering genes with co-varying expression. However, the methods, algorithms, data analysis systems and computer program are also useful for discovering other co-varying biological variables.

As will be appreciated by one of skill in the art, the present invention may be embodied as a method, data processing system or computer software program products.

Accordingly, the present invention may take the form of data analysis systems, methods, analysis software and etc. Software written according to the present invention is to be stored in some form of computer readable medium, such as memory, or CD-ROM. The software of the invention may be transmitted over a network and executed by a processor in a remote

5   location. The software may also be embedded in the computer readable medium of a hardware, such as an analytical instrument.

Fig. 1 illustrates an example of a computer system that may be used to execute the software of an embodiment of the invention. Fig. 1 shows a computer system 1 that includes a display 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one

10   or more buttons for interacting with a graphic user interface. Cabinet 7 houses a CD-ROM or DVD-ROM drive 13, system memory and a hard drive (see, Fig. 2) which may be utilized to store and retrieve software programs incorporating computer code that implements the invention, data for use with the invention and the like. Although a CD 15 is shown as an exemplary computer readable medium, other computer readable storage media including

15   floppy disk, tape, flash memory, system memory, and hard drive may be utilized. Additionally, a data signal embodied in a carrier wave (e.g., in a network including the internet) may be the computer readable storage medium.

Fig. 2 shows a system block diagram of computer system 1 used to execute the software of an embodiment of the invention. As in Fig. 1, computer system 1 includes

20   monitor 3, and keyboard 9, and mouse 11. Computer system 1 further includes subsystems such as a central processor 51, system memory 53, fixed storage 55 (e.g., hard drive), removable storage 57 (e.g., CD-ROM), display adapter 59, sound card 61, speakers 63, and network interface 65. Other computer systems suitable for use with the invention may include additional or fewer subsystems. For example, another computer system may include

25   more than one processor 51 or a cache memory. Computer systems suitable for use with the invention may also be embedded in a measurement instrument. The embedded systems may control the operation of, for example, a GeneChip® Probe array scanner as well as executing computer codes of the invention.

30   I.   Gene Expression Profile

The present invention provides methods, data analysis systems and computer software (program) products suitable for analyzing biological profiles, particularly large scale gene expression profiles. A "biological profile", as used herein, generally refers to a

collection variables reflecting the state of a biological sample at a given time in a given environment. A "biological sample," as used herein, generally refers to any biological materials such as a cell, a collection of cultured cells, a tissue sample, a biopsy, and an organism. For example, the state of cultured human fibroblast cells treated with a drug for 3

5   hours may be assessed by measuring the expression of a number of genes of the human fibroblast cells. The gene expression values are biological variables that reflect the state of the human fibroblast cells. The collection of the gene expression values constitutes a gene expression profile. For illustration purpose, the invention will be described using exemplary embodiments that analyze gene expression profiles. However, it would be apparent to those

10   skilled in the art, the invention is also useful for analyzing other biological profiles.

In some instances, a gene expression profile may be represented by the following equation:

$$G = [g_1, g_2, \ldots g_i, g_n];$$

Where: G is a gene expression profile;

15   g_1 - g_i are gene expression values, such as transcript levels, of
                       individual genes; and

                       n is the number of genes measured.

As used herein, the "dimension" or "Dimensionality" of a gene expression profile is generally referred to the number of genes in the profile (n). More generally, when

20   applied to a profile consisting of a number of biological variables, the term "dimension" or dimensionality of the profile is referred to the number of variables in the profile. For example, a GeneChip ® HuF1 array (Affymetrix, Inc., Santa Clara, CA) measures the expression of 6,000 human genes in a biological sample. The result of such a measurement is a gene expression profile with a dimension (n) of 6,000.

25   In some preferred embodiments, a high density oligonucleotide array is used to hybridize with a target nucleic acid sample to detect the expression level of a large number of genes, preferrably more than 10, more preferably more than 100, even more preferably more than 1000 genes and most preferably more than 2000 genes. Activity of a gene is reflected by the activity of its product: the proteins or other molecules encoded by the gene and that

30   perform biological functions. Measuring the activity of a gene product is, however, often difficult. Instead, the immunological activities or the amount of the final product or its peptide processing intermediates are determined as a measure of the gene activity. More frequently, the amount or activity of intermediates, such as transcripts, RNA processing

intermediates, mature mRNAs are detected as a measurement of gene activity. In many cases, the form and function of the final products of a gene is unknown. In those cases, the activity of gene is measured only by the amount or activity of transcripts, RNA processing intermediates, mature mRNAs.

Any methods that measure the activity of a gene are useful for at least some embodiments of this invention. For example, traditional Northern blotting and hybridization, nuclease protection, and RT- PCR have been used for detecting gene activity.

In some preferred embodiments, massive parallel gene expression monitoring methods are employed. Such methods are described in, for example, U.S. Provisional Application Serial Number 60/035,327, filed on 1/13/1997, PCT Application Number PCT/US/98/01206, filed on 1/12/ 98; Lockhart *et al.*, 1996, Nature Biotechnology. 14:1674-1680; Wodicka *et al.*, 1997, Nature Biotechnology. 15:1359, all incorporated in their entities by reference for all purposes. In such embodiments, a gene expression profile may contain level of transcripts of a plurality of genes. Alternatively, a gene expression profile may comprise of a plurality of measurements that are correlated to the level of transcripts of a plurality of genes. In some other embodiments, a gene expression profile contain the change of the levels of transcripts of a plurality of genes. For example, in one particularly preferred embodiment, a gene expression profile contains the fold change of transcripts of a plurality of genes during a developmental process. One of skill in the art would appreciate that gene expression profiles may contain other variables reflect the dynamics of the expression of a plurality of genes. For example, in one particularly preferred embodiment, a gene expression profile may contain Intensity values for a plurality of genes. Gene expression monitoring methods are also described in Section III, *infra*.

II.   Reduction of Dimensionality of Gene Expression Profiles

As discussed above, a gene expression profile may typically contain a large number of variables. For example, a GeneChip ® HuGeneFL array (Affymetrix, Inc., Santa Clara, CA) contains probes interrogating approximately 5,600 full-length human genes from UniGene(Build 18), GeneBank, and TIGR databases. A gene expression profile obtained using the HuGeneFL array will contain values for variables reflecting the expression of approximately 6000 genes.

The massive parallel gene expression monitoring technology has been used to study the relationship among the large number of genes or the response of those genes to

various treatments. Typically, a biological sample is subjected to various treatments (or several similar biological samples, each of which is subject to one treatment) and gene expression profiles of the biological sample under the various treatments are collected for further statistical analysis.

5   One of skill in the art would appreciate that before several gene expression profiles can be meaningfully compared, normalization must be performed to adjust experimental and other variations. Methods for such normalization is described by exemplary embodiments in Section IV, *infra*.

          An exemplary collection of gene expression profiles may be represented by the

10   following matrix:

$$
\begin{pmatrix} G_1 \\ \cdot \\ G_i \\ \cdot \\ G_m \end{pmatrix} = \begin{pmatrix} g_{1,1} & \cdot & \cdot & \cdot & g_{n,m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & g_{i,j} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{1,m} & \cdot & \cdot & \cdot & g_{n,m} \end{pmatrix}
$$

          wherein: G_j is the *j*th gene expression profile;

                                 $g_{i,j}$ is the expression value of *i*th gene in the *j* gene

15   expression profile; and

                                 n is the number of genes in the gene expression profiles; and

                                 m is the number of gene expression profiles.

          Typically, the number of gene expression profiles (*m*) is generally much

20   smaller than the number (*n*) of genes measured. For example, in a study of the action of three drug candidates on the genes of human cells, cells may be treated with three drug candidate and a control substance. If each treatment is repeated three times, a total of 3 x 4 = 12 gene expression profiles may be generated (*m*=12). In such a study, the expression of approximately 6000 genes (*n*=6000) may be monitored.

25   Because of the large number of variables, *i.e.*, a high dimensionality (high *n*), and the relatively low number of measurements per gene (low *m*), direct application of certain standard statistical techniques, such as cluster analysis, may not produce meaningful result. The problem is worsening when the treatments are non-toxic levels of drugs or drug candidates. Under such treatments, the changes in gene expression level may be small,

comparing with the changes during a development process or under toxic level of drug treatments. The small ranges of variation for each variable make it more difficult to identify patterns with statistical significance.

Accordingly, the present invention provides methods for reducing the
5   dimension (n) of the gene expression profiles without the loss of significant amount of information in the gene expression profiles. The reduction of dimension (n) of certain gene expression profiles, preferably to smaller than 1000, more preferably to smaller than 100, and most preferably to smaller than 50, makes it possible to meaningfully perform certain statistical analyses, such as the cluster analysis.   In addition, this invention also provide
10 effective data visualization methods for better appreciation of the interaction between genes and environmental factors (such as drug treatment).

Fig. 3 illustrates a process for gene expression profile analysis, preferably in a computer system. Gene expression profiles (1) are inputted into the computer system. Computer codes for reducing the dimension of the gene expression profiles are executed (2)
15 to obtain gene expression profiles with reduced dimension (often referred to as "reduced profile", "reduced biological profiles", "reduced gene expression profiles" etc.(3)). Computer code for cluster analysis or other statistical analysis is then executed on the reduced profiles to identify patterns. Fig. 4 illustrates an iterative process for reducing the dimension of gene expression profiles. The process is also preferably performed in a computer system.
20 Gene expression profiles (1) are inputted into the computer system. A dimension reduction (2) is performed on the gene expression profiles (1) by executing code for dimension reduction. Computer code for cluster analysis (4) is executed on the reduced gene expression profiles (3). The result of the cluster analysis is evaluated (6). If the result is determined to be acceptable for pattern recognition (5), the process stops. However, if the result is
25 determined to be unacceptable for pattern recognition (7), the reduced gene expression profiles are subjected to additional reduction of dimension. The process (described below) for reducing dimension is preferably different for different round of reduction. In some embodiments, computer program products are provided to perform these processes in a computer system. The computer program products contain code described above and these
30 codes are stored in suitable computer readable media.

In some embodiments, the dimension of a gene expression profiles is reduced by removing expression values for genes that are not significantly detected in any profile. In one preferred embodiment, the gene expression profiles are obtained using the

GeneChip® Probe Array. In the preferred embodiment, expression value for genes that are not detected in any of the profiles are removed from further analysis. Whether or not the expression of a gene is detectable may be determined by an absolute call decision matrix (decision matrix used for analyzing GeneChip ® Probe Array measurements are discussed in

5   Section IV, *infra*) . Data mining software such as those described in Balaban *et al.*, 1999, Method and Apparatus for Providing an Expression Data Mining Database, U.S. Patent Application (Attorney Docket Number 018547-033841US) filed on July 15, 1999 may be useful for determining whether a gene is expressed in a sample.

Fig. 5 illustrates a process for reduction of dimension of gene expression

10   profiles. The process is preferably performed in a computer system. Gene expression profiles (1) are inputted into a computer system. The expression of individual genes are analyzed (2). The expression of the gene is evaluated (3). If the expression is above a threshold value in any of the profiles, the gene is kept in the profiles. However, if the gene is not expressed above a threshold value in any of the profiles, the gene is removed from the

15   profiles (4). The process is repeated until all genes are evaluated (5). Then, the reduced gene expression profiles are outputted or used for statistical analysis. The process is accomplished by executing computer codes.

In some other embodiments, the dimensionality of a gene expression profile that reflects the change of gene expression between treatments is reduced by isolating the

20   genes which has the highest expression levels across treatments. In some preferred embodiments, the dimension (or number of variables) in an expression profile is reduced by selecting no more than 1000, preferably no more than 500, more preferable no more than 100 genes with the highest expression level across treatments. In one particularly preferred embodiment using the GeneChip® Probe array, genes for which the Average Difference

25   Intensity of Experiments (A.D.I.E.) are selected. In this embodiment, a plurality of probes are used to measure the expression of a gene. As explained in more details in section IV, *infra*, the measurement Average Difference Intensity may be calculated based upon the hybridization of the transcript of a gene and its corresponding probe set. A.D.I.E. of a gene is the mean of the Average Difference Intensities of the gene in all the treatments in an

30   experiment.

In some other embodiments, genes whose fold change has the highest variance across the different treatment are isolated for cluster analysis. In some additional

embodiments, genes whose fold change variable behaves most differently among treatments are isolated for cluster analysis.

In another aspect of the invention, the methods of studying drugs and/or drug candidates are provided. As used herein, the term "drug candidate" is generally referred to

5   any chemical compound or composition. The term "drug" is generally referred to any chemical compound or composition that has been shown to possess the ability to produce a desired biochemical or physiological change in a target.

In some embodiments of the invention, a biological sample is treated with drug and/or drug candidates. The gene expression of the biological sample is monitored to

10   obtain a set of gene expression profiles. At least one of the gene expression profile reflects the state of the biological sample under one of the drug and/or drug candidates. The dimension of the expression profiles is reduced according to the methods of the invention to obtain a reduced set of expression profiles. A cluster analysis is performed on the set of reduced gene expression profiles to cluster the drug and/or drug candidates. A short distance

15   between two drugs may indicate a similar mode of action for the two drugs.

In some other embodiments, drugs with known mode of action are compared with a drug candidate whose mode of action is unknown. Gene expression profiles reflecting the action of the drugs and the drug candidate may be subjected to reduction of dimension according to the methods of the invention to obtain a set of reduced gene expression profiles.

20   A cluster analysis is performed on the reduced gene expression profiles to obtain distances between the drugs and the drug candidate. The mode of action of the drug candidate is likely to be similar to that of the drug whose distance from the drug candidate is the smallest. Similarly, the methods of the invention may be used to predict the activity of a drug candidate. In such embodiments, the activities of the drug candidate is most likely to be

25   similar to these of the drug whose distance from the drug candidate is the smallest.

III.   Statistical Analysis of Reduced Gene Expression Profiles

In some preferred embodiments, reduced gene expression profiles are analyzed using statistical method to identify patterns and fingerprints. One particularly preferred

30   method is cluster analysis. Cluster analysis is an effective means to organize and explore relationships in data. Generally, there are two cluster approaches:

1. Cluster genes. In this case, the variables are the treatments (such as a time series, or drug concentration series) or subjects (such as cancerous vs. health tissues);

2. Cluster treatments. In this case, the variables are the reduced set of genes.

A step of normalization is generally included for cluster analysis protocols. Normalization allows patterns of expression to be compared independent of absolute expression levels.

In some embodiments, the normalization may be performed by normalizing the values to the range of data. Scaling may also be used for normalization purpose.

One of skill in the art would appreciate that there is not a single cluster analysis method that is most appropriate for use in all situations. Specific procedures must be selected with the consideration of the advantages and disadvantages of the specific procedures and the characteristics of the data.

Two types of clustering, k-means clustering or self organizing maps and hierarchical clustering, are particularly useful for use with methods of the invention.   k-means clustering is a partitioning method that constructs k clusters. K is generally fixed and supplied by a user. An object (such as a gene or a treatment) can only belong to one cluster. K-mean clustering or self organizing map has the advantages that points are re-evaluated and errors do not propagate. The disadvantages include the need to know the number of clusters, assumption that the clusters are round and assumption that the clusters are the same size.

There are two approaches to hierarchical systems. Agglomerative technique is fast and the number of the clusters need not to be known in advance. Disadvantages of the agglomerative techniques include early error progagation, no re-evaluation of a member, tendency to join close clusters and all members must be joined. Divisive methods have the advantages that k does not need to be known in advance and all members need not be joined. Disadvantages include early error progagation, no re-evaluation of members and low speed. Additional information about cluster analysis may be found in, for example, Tamayo, *et al.*, 1999, Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation, Proc. Natl. Acad. Sci. USA 96:2907-12; Eisen *et al.*, 1998, Cluster Analysis and Display of Genome-Wide Expression Patterns, Proc. Natl. Acad. Sci. USA 95:14863-1486; Wen *et al.*, 1998, Large Scale Temporal Gene Expression Mapping of Central Nervous System Development, Proc. Natl. Acad. Sci. USA 95:334-339; Alon *et al.*, Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor

and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proc. Natl. Acad. Sci. USA 96:6745-6750, all incorporated herein by reference for all purposes.

As the example (Section V, *infra*) shows, the methods of the invention greatly enhanced the application of cluster analysis to identify the patterns and fingerprints in gene expression profiles. In the example, cells were treated with several types of drugs. Gene expression profiles obtained from the cells were subjected to cluster analysis. As the example shows, direct application of the cluster analysis is overwhelmed by noises. After reduction of the dimension of the gene expression profiles, known drug effects and properties, such as activation mode and target specificity, can be correlated with expression patterns that are likely to be the signature for a specific drug effect or property.

In one aspect of the invention, method and computer program products for visualization are provided. In some embodiments, principal components analysis is used for data analysis and visualization. In this technique, an ordered set of orthogonal variables is constructed that explain the most variation in the data with the fewest number of variables. The single linear combination of the variables that has the highest variance is the first principal component. The highest variance linear combination orthogonal to the first principal component is the second principal component, and so forth. In some embodiments, the first two principal components are computed and graphed. As the example (Fig. 8) shows, principal components analysis facilitates the detection of patterns in the data, including the separation that permitted clustering for some cases, and to see the collapse of data that made separation difficult for other subsets.

Some embodiments of the computer program product of the invention contains computer codes for displaying data using the principal component method of the invention. The computer program product comprises: a) computer code that receives a plurality of biological profiles, each of the profiles comprises a plurality of biological variables; b) computer code that reduces the number of the biological variables to obtain a set of reduced profiles; c) computer code that performs a principal component analysis of the reduced profiles; d) computer code that displays the biological variables according a first axis and a second axis, wherein the first axis is the first component of the principal component analysis and the second axis is the second component of the principal component analysis, and wherein the first axis is perpendicular to the second axis; and e) computer readable medium that stores the computer codes.

IV.   Gene Expression Monitoring Methods

As discussed above, any methods that measure the activity of a gene are useful for at least some embodiments of this invention. For example, traditional Northern blotting and hybridization, nuclease protection, RT- PCR and differential display have been used for
5   detecting gene activity. Those methods are useful for some embodiments of the invention. However, this invention is most useful in conjunction with methods for detecting the expression of a large number of genes.

High density arrays are particularly useful for monitoring the expression control at the transcriptional, RNA processing and degradation level. The fabrication and
10   application of high density arrays in gene expression monitoring have been disclosed previously in, for example, U.S. Patent No. 5,800, 992, U.S. Application Ser. No. 08/772,376 (attorney docket number 2013.2), all incorporated herein for all purposed by reference. In some embodiment using high density arrays, high density oligonucleotide arrays are synthesized using methods such as the Very Large Scale Immobilized Polymer Synthesis
15   (VLSIPS) disclosed in U.S. Pat. No. 5,445,934 incorporated herein for all purposes by reference. Each oligonucleotide occupies a known location on a substrate. A nucleic acid target sample is hybridized with a high density array of oligonucleotides and then the amount of target nucleic acids hybridized to each probe in the array is quantified. One preferred quantifying method is to use confocal microscope and fluorescent labels. The GeneChip®
20   Probe Array system (Affymetrix, Santa Clara, CA) is particularly suitable for quantifying the hybridization; however, it is apparent to those of skill in the art that any similar systems or other effectively equivalent detection methods can also be used.

High density arrays are suitable for quantifying a small variations in expression levels of a gene in the presence of a large population of heterogeneous nucleic
25   acids. Such high density arrays can be fabricated either by de novo synthesis on a substrate or by spotting or transporting nature nucleic acid sequences onto specific locations of substrate. Nucleic acids are purified and/or isolated from biological materials, such as a bacteria plasmid containing a cloned segment of sequence of interest. Suitable nucleic acids are also produced by amplification of templates. As a nonlimiting illustration, polymerase
30   chain reaction, and/or in vitro transcription, are suitable nucleic acid amplification methods.

Synthesized oligonucleotide arrays are particularly preferred for this invention. Oligonucleotide arrays have numerous advantages, as opposed to other methods, such as

efficiency of production, reduced intra- and inter array variability, increased information content and high signal to noise ratio.

Preferred high density arrays for gene function identification and genetic network mapping comprise greater than about 100, preferably greater than about 1000, more

5   preferably greater than about 16,000 and most preferably greater than 65,000 or 250,000 or even greater than about 1,000,000 different oligonucleotide probes, preferably in less than 1 cm^2 of surface area. The oligonucleotide probes range from about 5 to about 50 or about 500 nucleotides, more preferably from about 10 to about 40 nucleotide and most preferably from about 15 to about 40 nucleotides in length.

10

A.   Massive Parallel Gene Expression Monitoring

One preferred method for massive parallel gene expression monitoring is based upon high density nucleic acid arrays.

Generally those methods of monitoring gene expression involve (a) providing

15   a pool of target nucleic acids comprising RNA transcript(s) of one or more target gene(s), or nucleic acids derived from the RNA transcript(s); (b) hybridizing the nucleic acid sample to a high density array of probes and (c) detecting the hybridized nucleic acids and calculating a relative and/or absolute expression (transcription, RNA processing or degradation) level.

(A).   Providing a Nucleic Acid Sample

20   One of skill in the art will appreciate that it is desirable to have nucleic samples containing target nucleic acid sequences that reflect the transcripts of interest. Therefore, suitable nucleic acid samples may contain transcripts of interest.   Suitable nucleic acid samples, however, may contain nucleic acids derived from the transcripts of interest. As used herein, a nucleic acid derived from a transcript refers to a nucleic acid for whose

25   synthesis the mRNA transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from a transcript, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, etc., are all derived from the transcript and detection of such derived products is indicative of the presence and/or abundance of the original transcript in a sample. Thus, suitable samples

30   include, but are not limited to, transcripts of the gene or genes, cDNA reverse transcribed from the transcript, cRNA transcribed from the cDNA, DNA amplified from the genes, RNA transcribed from amplified DNA, and the like.   Transcripts, as used herein, may include, but not limited to pre-mRNA nascent transcript(s), transcript processing intermediates,

mature mRNA(s) and degradation products.   It is not necessary to monitor all types of transcripts to practice this invention. For example, one may choose to practice the invention to measure the mature mRNA levels only.

In one embodiment, such sample is a homogenate of cells or tissues or other

5   biological samples. Preferably, such sample is a total RNA preparation of a biological sample. More preferably in some embodiments, such a nucleic acid sample is the total mRNA isolated from a biological sample. Those of skill in the art will appreciate that the total mRNA prepared with most methods includes not only the mature mRNA, but also the RNA processing intermediates and nascent pre-mRNA transcripts. For example, total mRNA

10   purified with poly (T) column contains RNA molecules with poly (A) tails. Those poly A+ RNA molecules could be mature mRNA, RNA processing intermediates, nascent transcripts or degradation intermediates.

Biological samples may be of any biological tissue or fluid or cells.
Frequently the sample will be a "clinical sample" which is a sample derived from a patient.

15   Clinical samples provide rich source of information regarding the various states of genetic network or gene expression. Some embodiments of the invention are employed to detect mutations and to identify the function of mutations. Such embodiments have extensive applications in clinical diagnostics and clinical studies. Typical clinical samples include, but are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy

20   samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

Another typical source of biological samples are cell cultures where gene expression states can be manipulated to explore the relationship among genes. In one aspect of the invention, methods are provided to generate biological samples reflecting a wide

25   variety of states of the genetic network.

One of skill in the art would appreciate that it is desirable to inhibit or destroy RNase present in homogenates before homogenates can be used for hybridization. Methods of inhibiting or destroying nucleases are well known in the art. In some preferred embodiments, cells or tissues are homogenized in the presence of chaotropic agents to inhibit

30   nuclease. In some other embodiments, RNase are inhibited or destroyed by heart treatment followed by proteinase treatment.

Methods of isolating total mRNA are also well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in

detail in Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993) and Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and

5   Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993)).

In a preferred embodiment, the total RNA is isolated from a given sample using, for example, an acid guanidinium-phenol-chloroform extraction method and polyA+ mRNA is isolated by oligo dT column chromatography or by using (dT)n magnetic beads (see, e.g., Sambrook et al., Molecular Cloning: A Laboratory Manual (2nd ed.), Vols. 1-3,

10   Cold Spring Harbor Laboratory, (1989), or Current Protocols in Molecular Biology, F. Ausubel et al., ed. Greene Publishing and Wiley-Interscience, New York (1987)).

Frequently, it is desirable to amplify the nucleic acid sample prior to hybridization. One of skill in the art will appreciate that whatever amplification method is used, if a quantitative result is desired, care must be taken to use a method that maintains or

15   controls for the relative frequencies of the amplified nucleic acids to achieve quantitative amplification.

Methods of "quantitative" amplification are well known to those of skill in the art. For example, quantitative PCR involves simultaneously co-amplifying a known quantity of a control sequence using the same primers. This provides an internal standard that may be

20   used to calibrate the PCR reaction. The high density array may then include probes specific to the internal standard for quantification of the amplified nucleic acid.

Other suitable amplification methods include, but are not limited to polymerase chain reaction (PCR) (Innis, et al., PCR Protocols. A guide to Methods and Application. Academic Press, Inc. San Diego, (1990)), ligase chain reaction (LCR) (see Wu

25   and Wallace, Genomics, 4: 560 (1989), Landegren, et al., Science, 241: 1077 (1988) and Barringer, et al., Gene, 89: 117 (1990), transcription amplification (Kwoh, et al., Proc. Natl. Acad. Sci. USA, 86: 1173 (1989)), and self-sustained sequence replication (Guatelli, et al., Proc. Nat. Acad. Sci. USA, 87: 1874 (1990)).

Cell lysates or tissue homogenates often contain a number of inhibitors of

30   polymerase activity. Therefore, RT-PCR typically incorporates preliminary steps to isolate total RNA or mRNA for subsequent use as an amplification template. One tube mRNA capture method may be used to prepare poly(A)+ RNA samples suitable for immediate RT-

PCR in the same tube (Boehringer Mannheim). The captured mRNA can be directly subjected to RT-PCR by adding a reverse transcription mix and, subsequently, a PCR mix.

In a particularly preferred embodiment, the sample mRNA is reverse transcribed with a reverse transcriptase and a primer consisting of oligo dT and a sequence

5   encoding the phage T7 promoter to provide single stranded DNA template. The second DNA strand is polymerized using a DNA polymerase. After synthesis of double-stranded cDNA, T7 RNA polymerase is added and RNA is transcribed from the cDNA template. Successive rounds of transcription from each single cDNA template results in amplified RNA. Methods of in vitro polymerization are well known to those of skill in the art (see, e.g., Sambrook,

10   supra.) and this particular method is described in detail by Van Gelder, et al., Proc. Natl. Acad. Sci. USA, 87: 1663-1667 (1990). Moreover, Eberwine et al. Proc. Natl. Acad. Sci. USA, 89: 3010-3014 provide a protocol that uses two rounds of amplification via in vitro transcription to achieve greater than 10^6 fold amplification of the original starting material thereby permitting expression monitoring even where biological samples are limited.

15   CRNA amplification methods disclosed in U.S. Provisional Application No. , , attorney docket number 3283, files on December 9, 1999.

It will be appreciated by one of skill in the art that the direct transcription method described above provides an antisense (aRNA) pool. Where antisense RNA is used as the target nucleic acid, the oligonucleotide probes provided in the array are chosen to be

20   complementary to subsequences of the antisense nucleic acids. Conversely, where the target nucleic acid pool is a pool of sense nucleic acids, the oligonucleotide probes are selected to be complementary to subsequences of the sense nucleic acids. Finally, where the nucleic acid pool is double stranded, the probes may be of either sense as the target nucleic acids include both sense and antisense strands.

25   The protocols cited above include methods of generating pools of either sense or antisense nucleic acids. Indeed, one approach can be used to generate either sense or antisense nucleic acids as desired. For example, the cDNA can be directionally cloned into a vector (e.g., Stratagene's p Bluscript II KS (+) phagemid) such that it is flanked by the T3 and T7 promoters. In vitro transcription with the T3 polymerase will produce RNA of one sense

30   (the sense depending on the orientation of the insert), while in vitro transcription with the T7 polymerase will produce RNA having the opposite sense. Other suitable cloning systems

include phage lambda vectors designed for Cre-loxP plasmid subcloning (see e.g., Palazzolo et al., Gene, 88: 25-36 (1990)).

(B)   Hybridizing nucleic acids to high density array

1. Probe design

5   One of skill in the art will appreciate that an enormous number of array designs are suitable for the practice of this invention. The high density array will typically include a number of probes that specifically hybridize to the sequences of interest. In addition, in a preferred embodiment, the array will include one or more control probes.

The high density array chip includes "test probes." Test probes could be 10   oligonucleotides that range from about 5 to about 45 or 5 to about 500 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In other particularly preferred embodiments the probes are 20 or 25 nucleotides in length. In another preferred embodiments, test probes are double or single strand DNA sequences. DNA sequences are isolated or cloned from nature sources or 15   amplified from nature sources using nature nucleic acid as templates. These probes have sequences complementary to particular subsequences of the genes whose expression they are designed to detect. Thus, the test probes are capable of specifically hybridizing to the target nucleic acid they are to detect.

In addition to test probes that bind the target nucleic acid(s) of interest, the 20   high density array can contain a number of control probes. The control probes fall into three categories referred to herein as 1) Normalization controls; 2) Expression level controls; and 3) Mismatch controls which are designed to contain at least one base that is different from that of a target sequence. Normalization controls are oligonucleotide or other nucleic acid probes that are complementary to labeled reference oligonucleotides or other nucleic acid 25   sequences that are added to the nucleic acid sample. The signals obtained from the normalization controls after hybridization provide a control for variations in hybridization conditions, label intensity, "reading" efficiency and other factors that may cause the signal of a perfect hybridization to vary between arrays. In a preferred embodiment, signals (e.g., fluorescence intensity) read from all other probes in the array are divided by the signal (e.g., 30   fluorescence intensity) from the control probes thereby normalizing the measurements.

Virtually any probe may serve as a normalization control. However, it is recognized that hybridization efficiency varies with base composition and probe length. Preferred normalization probes are selected to reflect the average length of the other probes

20

present in the array, however, they can be selected to cover a range of lengths. The normalization control(s) can also be selected to reflect the (average) base composition of the other probes in the array, however in a preferred embodiment, only one or a few normalization probes are used and they are selected such that they hybridize well (i.e. no

5   secondary structure) and do not match any target-specific probes.

Expression level controls are probes that hybridize specifically with constitutively expressed genes in the biological sample. Virtually any constitutively expressed gene provides a suitable target for expression level controls. Typically expression level control probes have sequences complementary to subsequences of constitutively

10   expressed "housekeeping genes" including, but not limited to the ß-actin gene, the transferrin receptor gene, the GAPDH gene, and the like. Mismatch controls may also be provided for the probes to the target genes, for expression level controls or for normalization controls. Mismatch controls are oligonucleotide probes or other nucleic acid probes designed to be identical to their corresponding test, target or control probes except for the presence of one or

15   more mismatched bases. A mismatched base is a base selected so that it is not complementary to the corresponding base in the target sequence to which the probe would otherwise specifically hybridize. One or more mismatches are selected such that under appropriate hybridization conditions (e.g. stringent conditions) the test or control probe would be expected to hybridize with its target sequence, but the mismatch probe would not

20   hybridize (or would hybridize to a significantly lesser extent). Preferred mismatch probes contain a central mismatch. Thus, for example, where a probe is a 20 mer, a corresponding mismatch probe will have the identical sequence except for a single base mismatch (e.g., substituting a G, a C or a T for an A) at any of positions 6 through 14 (the central mismatch).

Mismatch probes thus provide a control for non-specific binding or cross-

25   hybridization to a nucleic acid in the sample other than the target to which the probe is directed. Mismatch probes thus indicate whether a hybridization is specific or not. For example, if the target is present the perfect match probes should be consistently brighter than the mismatch probes. In addition, if all central mismatches are present, the mismatch probes can be used to detect a mutation. The difference in intensity between the perfect match and

30   the mismatch probe ($I(PM)-I(MM)$) provides a good measure of the concentration of the hybridized material.

The high density array may also include sample preparation/amplification control probes. These are probes that are complementary to subsequences of control genes

21

selected because they do not normally occur in the nucleic acids of the particular biological sample being assayed. Suitable sample preparation/amplification control probes include, for example, probes to bacterial genes (e.g., Bio B) where the sample in question is a biological from a eukaryote.

5   The RNA sample is then spiked with a known amount of the nucleic acid to which the sample preparation/amplification control probe is directed before processing. Quantification of the hybridization of the sample preparation/amplification control probe then provides a measure of alteration in the abundance of the nucleic acids caused by processing steps (e.g. PCR, reverse transcription, in vitro transcription, etc.).

10   In a preferred embodiment, oligonucleotide probes in the high density array are selected to bind specifically to the nucleic acid target to which they are directed with minimal non-specific binding or cross-hybridization under the particular hybridization conditions utilized. Because the high density arrays of this invention can contain in excess of 1,000,000 different probes, it is possible to provide every probe of a characteristic length that

15   binds to a particular nucleic acid sequence. Thus, for example, the high density array can contain every possible 20 mer sequence complementary to an IL-2 mRNA.

There, however, may exist 20 mer subsequences that are not unique to the IL-2 mRNA. Probes directed to these subsequences are expected to cross hybridize with occurrences of their complementary sequence in other regions of the sample genome.

20   Similarly, other probes simply may not hybridize effectively under the hybridization conditions (e.g., due to secondary structure, or interactions with the substrate or other probes). Thus, in a preferred embodiment, the probes that show such poor specificity or hybridization efficiency are identified and may not be included either in the high density array itself (e.g., during fabrication of the array) or in the post-hybridization data analysis.

25   In addition, in a preferred embodiment, expression monitoring arrays are used to identify the presence and expression (transcription) level of genes which are several hundred base pairs long. For most applications it would be useful to identify the presence, absence, or expression level of several thousand to one hundred thousand genes. Because the number of oligonucleotides per array is limited in a preferred embodiment, it is desired to

30   include only a limited set of probes specific to each gene whose expression is to be detected.

As disclosed in U.S. Application Ser. No. 08/772,376, probes as short as 15, 20, or 25 nucleotide are sufficient to hybridize to a subsequence of a gene and that, for most genes, there is a set of probes that performs well across a wide range of target nucleic acid

concentrations. In a preferred embodiment, it is desirable to choose a preferred or "optimum" subset of probes for each gene before synthesizing the high density array.

2. Forming High Density Arrays.

Methods of forming high density arrays of oligonucleotides, peptides and
5 other polymer sequences with a minimal number of synthetic steps are known. The oligonucleotide analogue array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling. See Pirrung et al., U.S. Patent No. 5,143,854 (see also PCT Application No. WO 90/15070) and Fodor et al., PCT Publication Nos. WO 92/10092 and WO 93/09668
10 and US Ser. No. 07/980,523 which disclose methods of forming vast arrays of peptides, oligonucleotides and other molecules using, for example, light-directed synthesis techniques. See also, Fodor et al., Science, 251, 767-77 (1991). These procedures for synthesis of polymer arrays are now referred to as VLSIPS™ procedures. Using the VLSIPS™ approach, one heterogeneous array of polymers is converted, through simultaneous coupling at a
15 number of reaction sites, into a different heterogeneous array. See, U.S. Application Serial Nos. 07/796,243 and 07/980,523.

The development of VLSIPS™ technology as described in the above-noted U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, is considered pioneering technology in the fields of combinatorial synthesis and screening of
20 combinatorial libraries. More recently, patent application Serial No. 08/082,937, filed June 25, 1993 describes methods for making arrays of oligonucleotide probes that can be used to check or determine a partial or complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific oligonucleotide sequence.

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on
25 a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, e.g., a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithogaphic mask is used selectively to expose functional groups which are then ready to react with incoming 5'-photoprotected
30 nucleoside phosphoramidites. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid

surface. Combinatorial synthesis of different oligonucleotide analogues at different locations on the array is determined by the pattern of illumination during synthesis and the order of addition of coupling reagents.

In the event that an oligonucleotide analogue with a polyamide backbone is
5   used in the VLSIPS™ procedure, it is generally inappropriate to use phosphoramidite chemistry to perform the synthetic steps, since the monomers do not attach to one another via a phosphate linkage. Instead, peptide synthetic methods are substituted. See, e.g., Pirrung et al. U.S. Pat. No. 5,143,854.

Peptide nucleic acids are commercially available from, e.g., Biosearch, Inc.
10   (Bedford, MA) which comprise a polyamide backbone and the bases found in naturally occurring nucleosides. Peptide nucleic acids are capable of binding to nucleic acids with high specificity, and are considered "oligonucleotide analogues" for purposes of this disclosure.

In addition to the foregoing, additional methods which can be used to generate an array of oligonucleotides on a single substrate are described in co-pending Applications
15   Ser. No. 07/980,523, filed November 20, 1992, and 07/796,243, filed November 22, 1991 and in PCT Publication No. WO 93/09668. In the methods disclosed in these applications, reagents are delivered to the substrate by either (1) flowing within a channel defined on predefined regions or (2) "spotting" on predefined regions or (3) through the use of photoresist. However, other approaches, as well as combinations of spotting and flowing,
20   may be employed. In each instance, certain activated regions of the substrate are mechanically separated from other regions when the monomer solutions are delivered to the various reaction sites.

A typical "flow channel" method applied to the compounds and libraries of the present invention can generally be described as follows. Diverse polymer sequences are
25   synthesized at selected regions of a substrate or solid support by forming flow channels on a surface of the substrate through which appropriate reagents flow or in which appropriate reagents are placed. For example, assume a monomer "A" is to be bound to the substrate in a first group of selected regions. If necessary, all or part of the surface of the substrate in all or a part of the selected regions is activated for binding by, for example, flowing appropriate
30   reagents through all or some of the channels, or by washing the entire substrate with appropriate reagents. After placement of a channel block on the surface of the substrate, a reagent having the monomer A flows through or is placed in all or some of the channel(s).

24

The channels provide fluid contact to the first selected regions, thereby binding the monomer A on the substrate directly or indirectly (via a spacer) in the first selected regions.

Thereafter, a monomer B is coupled to second selected regions, some of which may be included among the first selected regions. The second selected regions will be in

5   fluid contact with a second flow channel(s) through translation, rotation, or replacement of the channel block on the surface of the substrate; through opening or closing a selected valve; or through deposition of a layer of chemical or photoresist. If necessary, a step is performed for activating at least the second regions. Thereafter, the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations.

10   In this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, a monomer C can be flowed

15   through still other channels, etc. In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of

20   forming channels or otherwise protecting a portion of the surface of the substrate. For example, according to some embodiments, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further

25   prevented from passing outside of their designated flow paths.

High density nucleic acid arrays can be fabricated by depositing presynthezied or nature nucleic acids in predined positions. As disclosed in the U.S. Application Ser. No. and its parent applications, previously incorporated for all purposed, synthesized or nature nucleic acids are deposited on specific locations of a substrate by light directed targeting and

30   oligonucleotide directed targeting. Nucleic acids can also be directed to specific locations in much the same manner as the flow channel methods. For example, a nucleic acid A can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a nucleic acid B can be delivered to and reacted with a second group of

activated reaction regions. Nucleic acids are deposited in selected regions. Another embodiment uses a dispenser that moves from region to region to deposit nucleic acids in specific spots. Typical dispensers include a micropipette or capillary pin to deliver nucleic acid to the substrate and a robotic system to control the position of the micropipette with

5   respect to the substrate. In other embodiments, the dispenser includes a series of tubes, a manifold, an array of pipettes or capillary pins, or the like so that various reagents can be delivered to the reaction regions simultaneously.

3. Hybridization

Nucleic acid hybridization simply involves contacting a probe and target

10   nucleic acid under conditions where the probe and its complementary target can form stable hybrid duplexes through complementary base pairing. The nucleic acids that do not form hybrid duplexes are then washed away leaving the hybridized nucleic acids to be detected, typically through detection of an attached detectable label. It is generally recognized that nucleic acids are denatured by increasing the temperature or decreasing the salt concentration

15   of the buffer containing the nucleic acids. Under low stringency conditions (e.g., low temperature and/or high salt) hybrid duplexes (e.g., DNA:DNA, RNA:RNA, or RNA:DNA) will form even where the annealed sequences are not perfectly complementary. Thus specificity of hybridization is reduced at lower stringency. Conversely, at higher stringency (e.g., higher temperature or lower salt) successful hybridization requires fewer mismatches.

20   One of skill in the art will appreciate that hybridization conditions may be selected to provide any degree of stringency. In a preferred embodiment, hybridization is performed at low stringency in this case in 6X SSPE-T at 37 C (0.005% Triton X-100) to ensure hybridization and then subsequent washes are performed at higher stringency (e.g., 1 X SSPE-T at 37 C) to eliminate mismatched hybrid duplexes. Successive washes may be

25   performed at increasingly higher stringency (e.g., down to as low as 0.25 X SSPE-T at 37 C to 50 C) until a desired level of hybridization specificity is obtained. Stringency can also be increased by addition of agents such as formamide. Hybridization specificity may be evaluated by comparison of hybridization to the test probes with hybridization to the various controls that can be present (e.g., expression level control, normalization control, mismatch

30   controls, etc.).

In general, there is a tradeoff between hybridization specificity (stringency) and signal intensity. Thus, in a preferred embodiment, the wash is performed at the highest stringency that produces consistent results and that provides a signal intensity greater than

approximately 10% of the background intensity. Thus, in a preferred embodiment, the
hybridized array may be washed at successively higher stringency solutions and read between
each wash. Analysis of the data sets thus produced will reveal a wash stringency above
which the hybridization pattern is not appreciably altered and which provides adequate signal
5   for the particular oligonucleotide probes of interest.

In a preferred embodiment, background signal is reduced by the use of a
detergent (e.g., C-TAB) or a blocking reagent (e.g., sperm DNA, cot-1 DNA, etc.) during the
hybridization to reduce non-specific binding. In a particularly preferred embodiment, the
hybridization is performed in the presence of about 0.5 mg/ml DNA (e.g., herring sperm
10   DNA). The use of blocking agents in hybridization is well known to those of skill in the art
(see, e.g., Chapter 8 in P. Tijssen, supra.)

The stability of duplexes formed between RNAs or DNAs are generally in the
order of RNA:RNA > RNA:DNA > DNA:DNA, in solution. Long probes have better duplex
stability with a target, but poorer mismatch discrimination than shorter probes (mismatch
15   discrimination refers to the measured hybridization signal ratio between a perfect match
probe and a single base mismatch probe). Shorter probes (e.g., 8-mers) discriminate
mismatches very well, but the overall duplex stability is low.

Altering the thermal stability (T_m) of the duplex formed between the target and the probe using, e.g., known oligonucleotide analogues allows for optimization of duplex stability and mismatch discrimination. One useful aspect of altering the T_m arises from the fact that adenine-thymine (A-T) duplexes have a lower T_m than guanine-cytosine (G-C)

5   duplexes, due in part to the fact that the A-T duplexes have 2 hydrogen bonds per base-pair, while the G-C duplexes have 3 hydrogen bonds per base pair. In heterogeneous oligonucleotide arrays in which there is a non-uniform distribution of bases, it is not generally possible to optimize hybridization for each oligonucleotide probe simultaneously. Thus, in some embodiments, it is desirable to selectively destabilize G-C duplexes and/or to increase

10   the stability of A-T duplexes. This can be accomplished, e.g., by substituting guanine residues in the probes of an array which form G-C duplexes with hypoxanthine, or by substituting adenine residues in probes which form A-T duplexes with 2,6 diaminopurine or by using the salt tetramethyl ammonium chloride (TMACl) in place of NaCl.

Altered duplex stability conferred by using oligonucleotide analogue probes

15   can be ascertained by following, e.g., fluorescence signal intensity of oligonucleotide analogue arrays hybridized with a target oligonucleotide over time. The data allow optimization of specific hybridization conditions at, e.g., room temperature (for simplified diagnostic applications in the future).

Another way of verifying altered duplex stability is by following the signal

20   intensity generated upon hybridization with time. Previous experiments using DNA targets and DNA chips have shown that signal intensity increases with time, and that the more stable duplexes generate higher signal intensities faster than less stable duplexes. The signals reach a plateau or "saturate" after a certain amount of time due to all of the binding sites becoming occupied. These data allow for optimization of hybridization, and determination of the best

25   conditions at a specified temperature.

Methods of optimizing hybridization conditions are well known to those of skill in the art (see, e.g., Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993)).

(C) Signal Detection

30   In a preferred embodiment, the hybridized nucleic acids are detected by detecting one or more labels attached to the sample nucleic acids. The labels may be incorporated by any of a number of means well known to those of skill in the art. However, in a preferred embodiment, the label is simultaneously incorporated during the amplification

step in the preparation of the sample nucleic acids. Thus, for example, polymerase chain reaction (PCR) with labeled primers or labeled nucleotides will provide a labeled amplification product. In a preferred embodiment, transcription amplification, as described above, using a labeled nucleotide (e.g. fluorescein-labeled UTP and/or CTP) incorporates a

5   label into the transcribed nucleic acids.

Alternatively, a label may be added directly to the original nucleic acid sample (e.g., mRNA, polyA mRNA, cDNA, etc.) or to the amplification product after the amplification is completed. Means of attaching labels to nucleic acids are well known to those of skill in the art and include, for example nick translation or end-labeling (e.g. with a

10   labeled RNA) by kinasing of the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (e.g., a fluorophore).

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include biotin for

15   staining with labeled streptavidin conjugate, magnetic beads (e.g., Dynabeads^{TM}), fluorescent dyes (e.g., fluorescein, texas red, rhodamine, green fluorescent protein, and the like), radiolabels (e.g., ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P), enzymes (e.g., horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (e.g., polystyrene, polypropylene, latex, etc.)

20   beads. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

Means of detecting such labels are well known to those of skill in the art. Thus, for example, radiolabels may be detected using photographic film or scintillation counters, fluorescent markers may be detected using a photodetector to detect emitted light.

25   Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label. One particular preferred methods uses colloidal gold label that can be detected by measuring scattered light.

The label may be added to the target (sample) nucleic acid(s) prior to, or after the

30   hybridization. So called "direct labels" are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so called "indirect labels" are joined to the hybrid duplex after hybridization. Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the

29

hybridization. Thus, for example, the target nucleic acid may be biotinylated before the
hybridization. After hybridization, an aviden-conjugated fluorophore will bind the biotin
bearing hybrid duplexes providing a label that is easily detected. For a detailed review of
methods of labeling nucleic acids and detecting labeled hybridized nucleic acids see

5   Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With
Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993)).

Fluorescent labels are preferred and easily added during an in vitro
transcription reaction. In a preferred embodiment, fluorescein labeled UTP and CTP are
incorporated into the RNA produced in an in vitro transcription reaction as described above.

10   Means of detecting labeled target (sample) nucleic acids hybridized to the
probes of the high density array are known to those of skill in the art. Thus, for example,
where a colorimetric label is used, simple visualization of the label is sufficient. Where a
radioactive labeled probe is used, detection of the radiation (e.g. with photographic film or a
solid state detector) is sufficient.

15   In a preferred embodiment, however, the target nucleic acids are labeled with a
fluorescent label and the localization of the label on the probe array is accomplished with
fluorescent microscopy. The hybridized array is excited with a light source at the excitation
wavelength of the particular fluorescent label and the resulting fluorescence at the emission
wavelength is detected. In a particularly preferred embodiment, the excitation light source is

20   a laser appropriate for the excitation of the fluorescent label.

The confocal microscope may be automated with a computer-controlled stage
to automatically scan the entire high density array. Similarly, the microscope may be
equipped with a phototransducer (e.g., a photomultiplier, a solid state array, a CCD camera,
etc.) attached to an automated data acquisition system to automatically record the

25   fluorescence signal produced by hybridization to each oligonucleotide probe on the array.
Such automated systems are described at length in U.S. Patent No: 5,143,854, PCT
Application 20 92/10092, and copending U.S. Application Ser. No. 08/195,889 filed on
February 10, 1994. Use of laser illumination in conjunction with automated confocal
microscopy for signal detection permits detection at a resolution of better than about 100 μm,

30   more preferably better than about 50 μm, and most preferably better than about 25 μm.

One of skill in the art will appreciate that methods for evaluating the
hybridization results vary with the nature of the specific probe nucleic acids used as well as
the controls provided. In the simplest embodiment, simple quantification of the fluorescence

intensity for each probe is determined. This is accomplished simply by measuring probe signal strength at each location (representing a different probe) on the high density array (e.g., where the label is a fluorescent label, detection of the amount of florescence (intensity) produced by a fixed excitation illumination at each location on the array). Comparison of the

5   absolute intensities of an array hybridized to nucleic acids from a "test" sample with intensities produced by a "control" sample provides a measure of the relative expression of the nucleic acids that hybridize to each of the probes.

One of skill in the art, however, will appreciate that hybridization signals will vary in strength with efficiency of hybridization, the amount of label on the sample nucleic

10   acid and the amount of the particular nucleic acid in the sample. Typically nucleic acids present at very low levels (e.g., < 1pM) will show a very weak signal. At some low level of concentration, the signal becomes virtually indistinguishable from background. In evaluating the hybridization data, a threshold intensity value may be selected below which a signal is not counted as being essentially indistinguishable from background.

15   (D) Data Analysis

Intensity values are analyzed to generate gene expression values which constitutes a gene expression profile. A suitable embodiment of the data analysis methods and algorithms are described below using an exemplary embodiment.

In this embodiment, a computer program is used to calculate a variety of

20   metrics using the hybridization intensities measured by the scanner. Some metrics utilize intensity data from the entire probe array and are used for Background and Noise calculations. Other metrics compare the intensities of the sequence-specific Perfect Match (PM, which is designed to be a prefect match with the target sequence) probe cells with their control Mismatch (MM, which is designed to be mismatch against a target sequence) probe

25   cells for each probe set, and are then used by a decision matrix to determine if a transcript is Present (P), Marginal (M), or Absent (A; undetected). Because this analysis involves data from a single experiment to generate a single gene expression profile, this analysis is occasionally referred to as absolute analysis or absolute analysis algorithm.

The analysis begins by calculating an Average Intensity value for every probe

30   cell. Then, the Background is calculated and subtracted from the intensities of all probe cells. The Noise (Q) is also calculated by determining the degree of pixel to pixel variation within the same probe cells used to calculate the background. Noise is one of the criteria used to

determine the significance of intensity differences between PM probe cells and their MM controls.

Next, the numbers of Positive and Negative probe pairs are determined for every probe set. A positive probe pair is one in which the intensity of the sequence-specific (PM) probe cell is significantly higher than the intensity of the control (MM) probe cell. A Negative probe pair is one in which the intensity of the MM probe cell is significantly higher than the intensity of the PM probe cell. The numbers of Positive and Negative probe pairs are used to derive metrics that describe the performance of each probe set. These metrics are the Positive Fraction and the Pos / Neg Ratio.Two other metrics that describe the performance of each probe set, the Log Average Ratio (Log Avg Ratio) and the Average Difference (Avg Diff), are also calculated. These metrics use probe cell intensities directly rather than relying on the numbers of Positive and Negative probe pairs. The Log Avg Ratio is derived from the ratio of PM probe cell intensity to that of the control MM.

The Avg Diff for each probe set, an average of the differences between every PM probe cell and its control MM probe cell, is directly related to the level of expression of the transcript.

Finally, a "decision matrix" is employed to determine the presence or absence of each transcript (the Absolute Call). This is accomplished by examining three of the analysis metrics: the Positive Fraction, the Pos / Neg Ratio, and the Log Average Ratio. The Absolute Call is displayed in the data output of a file in analysis software along with all the Analysis Metrics for every transcript.

GeneChip ® probe arrays are scanned at high pixel resolution, resulting in a large amount of data. In the case of a higher density probe array, which contains probe cells that measure 24mm x 24mm, the array is scanned at a resolution of 3mm per pixel. This creates ~8 pixels x 8 pixels (on average) for every probe cell, or a total of ~64 pixels per probe cell. A single intensity value for every probe cell, representative of the hybridization level of its target, is derived as follows. The bordering pixels of the probe cell are excluded. The remaining pixel intensity distribution is calculated, and the intensity value associated with 75% of the distribution is used as the Average Intensity of the probe cell.

Background is a measurement of the signal intensity caused by auto-fluorescence of the array surface and nonspecific binding of target or stain molecules (SAPE). The calculation is done as follows:

1)   The array is divided into sectors (16 by default).

2)   The software ranks probe cells by fluorescence intensity, identifies the lowest 2% (by default) for each sector, and calculates their average. The resulting value is the sectors' Background.

3)   The sectors' background is subtracted from the average intensities of all probe cells within that sector.

Noise and Background are distinct phenomena and they are calculated separately by the software. Noise (Q) results from small variations in the digitized signal observed by the scanner as it samples the probe array's surface. The level of noise is calculated by the software, and then used as one of the criteria to determine the significance of differences between PM and MM probe cells, and differences in probe set intensities across two probe arrays.

Noise is measured by examining the pixel to pixel variations in signal intensities. As shown in the equation below, it is calculated using the standard deviations of pixel intensities of the background probe cells. If Normalization or Scaling is used in the analysis (discussed in a later section), the noise is scaled or normalized along with the rest of the data.

Absolute Analysis analyzes data from one probe array to determine if gene transcripts are detectable in the sample. This is accomplished by calculating a set of Absolute Metrics, the first of which are the number of Positive and Negative probe pairs.

When the intensity of the PM probe cell is significantly greater than that of the corresponding MM probe cell, the probe pair is termed Positive. When the intensity of the MM probe cell is significantly greater than that of the corresponding PM probe cell, the probe pair is termed Negative.

The significance is determined by calculating both the ratio (PM / MM) and the difference (PM -MM) associated with each probe pair. These values are then compared against two threshold values: the Statistical Difference Threshold (SDT) and the Statistical Ratio Threshold (SRT).

This is expressed mathematically as follows:

A probe pair is Positive if	A probe pair is Negative if:
(1) PM −MM ≥ SDT; and	(1) MM-PM ≥ SDT; and
(2) PM /MM ≥ SRT	(2) MM/PM ≥ SRT

Note: not all probe pairs will be scored as Positive or Negative.

The thresholds are defined as follows:

SDT is calculated by the software based on the noise, Q:

$$SDT = (Q) * (SDT\ mult)$$

SDT mult (SDT multiplier) is set by default to 2.0 when the single SAPE

5   staining protocol is used (usually with 50 μm feature arrays), and to 4.0 when the antibody

amplification protocol is used (usually with 20 μm feature arrays) SRT is set to 1.5 by

default.

The SDT mult and SRT can be modified by the user. Increasing

SRT and SDTmult makes the analysis more stringent and decreasing them makes the analysis

10   less stringent.

The numbers of Positive and Negative probe pairs, as well as PM and MM

intensities, are used to derive three additional Absolute Call Metrics for every transcript.

They are the Positive Fraction, the Pos / Neg Ratio, and the Log Avg Ratio. These metrics are

used to determine if a transcript is called "Present", "Marginal," or "Absent".

15   1)   Positive Fraction: a measure of the fraction of probe pairs in which the PM probe cells

have hybridized with a specific target to a greater level than the corresponding MM control. It

is calculated as the number of Positive probe pairs divided by the number of probe pairs used

in a probe set:

20   Positive Fraction = # positive probe pairs / # probe pairs used

Pairs Used is equal to the number of probe pairs in the set (usually 16 - 20) minus any that are

masked Pair

25   2)   Pos / Neg Ratio: the Ratio of Positive probe pairs to Negative probe pairs in a probe

set. It is calculated as follows:

Pos / Neg Ratio = #Positive probe pairs / # Negative probe pairs

30   3)   Log Avg Ratio: a metric that describes the hybridization performance of a probe set

by determining the ratio of the PM to MM intensities for each probe pair, taking the Log of

the resulting values, then averaging them across the probe set. Here is the equation:

$$Log\ Avg\ Ratio = 10 * [\Sigma \log (PM / MM)] / (Pairs\ in\ Avg)$$

Pairs in Avg: a "trimmed" probe set used in the Log Avg Ratio and Avg Difference calculations.

Pairs in Avg are determined as follows:

- When 8 PP or fewer are used: Pairs in Avg = Pairs Used

5   - When greater than 8 PP are used: perform "super scorings"

Superscoring: A process which filters probe pairs that are out of a given range when calculating Avg Diff and Log Avg Ration. The mean and standard deviation are calculated for intensity differences (PM-MM) across the entire probe set (excluding the highest and lowest

10   values), and values within a set number of standard deviations are not included in the calculation. The default value that defines that number of standard deviations is 3 (STP=3, user defined).

Each of the three metrics used to determine the Absolute Call (Pos/Neg Ratio, Positive Fraction, and Log Average Ratio) is weighted and entered into a decision matrix to

15   determine the status of a transcript. User-modifiable thresholds (called *min* and *max* for each of the three Absolute Call metrics) govern the way each metrics' value will influence the call for every transcript. Default values for these thresholds have been established through empirical testing.

In the software embodiment, the software product also performs additional

20   calculations on data from two separate probe array experiments in order to compare gene expression levels between two samples. This type of analysis is occasionally referred to as Comparison Analysis. This analysis employs Normalization or Scaling techniques to minimize differences in overall signal intensities between the two arrays allowing for more reliable detection of biologically relevant changes in the samples.

25   The Comparison Analysis begins with the user designating an Absolute Analysis of one probe array experiment as the source of Baseline data and a second probe array experiment as the source of Experimental data to be compared to the Baseline. The Experimental probe array is analyzed with the Absolute Algorithms, producing the Absolute Analysis results. The data is also analyzed using the Comparison Algorithms to derive

30   metrics that identify differences between the Experimental and Baseline probe arrays for every probe set. Some of the Comparison Analysis metrics are used in a decision matrix to derive a Difference Call, which indicates whether a transcript has Increased (I), Decreased (D), Marginally Increased (MI), Marginally Decreased (MD), or exhibits No Change (NC) in

expression level. In addition, a Fold Change calculation is carried out as an indication of the relative change of each transcript represented on the probe array.

Non-biological factors can contribute to the variability of data in many biological assays. In experiments, variations in the amount and quality of target hybridized
5   to the array, the amount of stain applied, or other experimental variables, may contribute to an overall variability in hybridization intensities. In order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized. Two mathematical techniques, similar in principle, are applied by the exemplary software product: Normalization and Scaling. These techniques can be applied by using data from a limited
10   (user-defined) group of probe sets, or from all probe sets (Global Normalization or Scaling). Here, the Global approach is explained in more detail.

Global Normalization is the computational technique in which the output of the experimental array is multiplied by a factor (the Normalization Factor, NF) to make its Average Intensity* equivalent to the Average Intensity of the baseline array.
15   In Global Scaling, the output of any experiment is multiplied by a factor (the Scaling Factor, SF) to make its Average Intensity equal to an arbitrary Target Intensity, set by the user. Scaling allows a number of experiments to become normalized to one Target Intensity, allowing comparison between any two experiments.

Comparison Analysis compares data from two probe arrays to determine
20   whether the expression level of each transcript has changed. This is accomplished by calculating a set of Comparison Metrics, the first of which are the Increase and Decrease probe pairs. These are defined as follows:

A probe pair is considered to Increase if the intensity difference between the PM and the MM probe cells in the experimental sample is significantly higher than in the
25   baseline sample. A probe pair is considered to Decrease if the intensity difference between the PM and the MM probe cells in the experimental sample is significantly lower than in the baseline sample.

Mathematically, two criteria must be met for a probe pair to show a significant Increase:
30   (1) (PM - MM) exp - (PM - MM) base $\geq$ Change Threshold (CT)

And

(2) [(PM - MM) exp - (PM - MM) base] / (PM - MM) base > Percent Change Threshold / 100

Likewise, two criteria must be met for a probe pair to show a significant decrease:

(1) (PM - MM) base - (PM - MM) exp > Change Threshold (CT)

And

5   (2) [(PM - MM) base - (PM - MM) exp] / (PM - MM) base > Percent Change Threshold / 100

Change Threshold (CT): Calculated by the software using the SDT (Statistical Difference Threshold; see *supra*) of both experimental and baseline data. Alternatively, the CT can be user-defined by entering a value for the CTmultiplier (CT mult), which is then

10   multiplied by the noise (Q) of the baseline or experimental data, whichever is greater.

Percent Change Threshold (PCT): defined by the user.

Four Comparison metrics are used to determine whether each transcript's expression level has changed between the baseline and experimental samples. They are:

1) Max (Increase / PP used, Decrease / PP used)

15   2) Increase/Decrease Ratio

3) Log Average Ratio Change

4) Dpos-Dneg Ratio (difference positive-difference negative)

1) Max (Increase / PP used, Decrease / PP used). This metric calculates the number of probe

20   pairs that have changed in a certain direction:

Increase / PP used = number of Increased Probe Pairs / number of probe pairs used

Decrease / PP used = number of Decreased Probe Pairs / number of probe pairs used. The larger of these values will be used in the decision matrix.

25   2) Increase/Decrease: Ratio of increased probe pairs over decreased probe pairs.

Increase / Decrease Ratio = # Increased Probe Pairs / # Decreased Probe Pairs.

3) Log Average Ratio Change

For every transcript, the difference between the Log Avg Ratio of the baseline and experimental data is calculated. The Log Avg Ratios are recomputed for each probe set based

30   on probe pairs used in both the baseline and experimental probe arrays (the recomputed values are not displayed by the software for the baseline data).

Log Avg Ratio Change = Log Avg exp - Log Avg base

4) Difference Positive - Difference Negative Ratio (Dpos-Dneg Ratio):

This metric combines the change in the number of Positive probe pairs and the change in the number of Negative probe pairs between the baseline and experimental data into one metric for every probe set. It is calculated as follows:

Dpos - Dneg Ratio = (Positive Change) - (Negative Change) / # of PP used

5   Where:

Positive Change = # Positive Probe Pairs exp - # Positive Probe Pairs base

Negative Change: # Negative Probe Pairs exp - # Negative Probe Pairs base

The two metrics above, Log Avg Ratio Change and Dpos - Dneg Ratio are typically positive when a transcript changes from a very low to a relatively high expression level, and

10   are typically negative if a transcript expression level changes from a relatively high to a very low or an undetectable level. If a transcript is present in both the baseline and experimental samples, these two metrics may be close to zero despite an increase or decrease in the level of the transcript.

The Difference Call Decision Matrix is an algorithm that generates one of five

15   outcomes for every transcript: Increase (I), Marginally Increase (MI), Decrease (D), Marginally Decrease (MD), and No Change (NC). The following four metrics are weighted differently and entered into the Decision Matrix:

1) Max [Increase / Total , Decrease / Total]

2) Increase / Decrease Ratio

20   3) Log Average Ratio Change

4) Dpos-Dneg Ratio

The Difference Call Decision Matrix relies on user-modifiable thresholds in much the same way as the Absolute Call Decision Matrix. The thresholds (called min and max for each of the four Difference Call metrics listed above) define the boundaries where

25   each metric may change the outcome of the Decision Matrix. Default values for these thresholds have been established at Affymetrix through extensive empirical testing.

Since the Avg Diff of a transcript is directly related to its expression level, an estimate of the Fold Change of the transcript between the baseline and experimental samples can be calculated. First, the Normalized or Scaled Avg Diff values are recomputed in both the

30   experimental and baseline data sets to include only probe pairs that are used in both the baseline and experimental arrays.

Next, the Avg Diff Change is determined as follows:

Avg Diff Change = Avg Diff exp - Avg Diff base

Then, the Fold Change is calculated with this equation:

$$FC = (\text{Avg Diff Change} / \max[\min(\text{Avg Diff exp}, \text{Avg Diff base}), Q M * Q c) +$$

$$\{+1 \text{ if Avg Diff exp} > \text{Avg Diff base} -1 \text{ if Avg Diff exp} < \text{Avg Diff base}\}$$

5   Where $Q c = \max$ (Q exp, Q base)

       And the   Q multiplier $Q M = 2.1$ in 50 μm feature arrays; 2.8 in 20 μm

                                           feature arrays.

This equation permits the expression of the Fold Change as a positive number when the transcript has increased over its baseline state, and as a negative number when the transcript

10   level declines.

If the Noise ($Q * Q M$) of either array is greater than the Avg Diff of the transcript (in either the baseline or experimental data), the Fold Change is calculated using the Noise. In this case the Fold Change is an approximation.

15   **IV.**   **Example**

This example illustrates one embodiment of methods of the invention and its application in studying drug activities.

A.   **Introduction**

Drug actions on a target cell is hypothesized to alter the mRNA expression

20   levels and phenotypic changes. Alteration of the mRNA levels may be correlated with drug target, activation mode, structure, efficacy or side effects. Some embodiments of the data analysis methods are used to analyze such correlations.

B.   **Experimental Protocols**

Human primary aortic smooth muscle cells, which are the target cells of many

25   α and β adrenergic compounds, were obtained from BioWhittaker and were cultured in SmGM media supplied by BioWhittaker in 60 mm or 100 mm tissue culture dish. Cells were treated with 39 compounds (Fig. 6 lists compounds and their properties) either alone or in combination at the concentration of 10^{-6} M for 8 hours in three separate experiments. These compounds include agonists and antagonists for α1, α2, β1, and β2 adrenergic receptors, as

30   well as angiotensin and dopamine targeted to different receptors. Total RNAs were isolated from cells before the treatment and from cells treated with these compounds and vehicle as a control using RNeasy kit supplied by QIAGEN. A total number of about 48 RNA samples were collected. These RNAs are analyzed twice in different experiment for mRNA

expression of more than 6000 human genes using Affymetrix HumF1Gene array according to the Gene Expression Technical Manual. A total of about 100 mRNA expression profiles were generated from the duplicate analysis of 48 RNA samples, 50 treatments with 39 compounds.

5   The scanned results of the hybridization were initially analyzed using GeneChip® Analysis Suite Version 3.0 (Affymetrix, Inc., Santa Clara, CA) to generate the basic analysis results, including the intensity of the expression, fold change of the expression levels, present call, and difference call.

          Subsequently, the data were filtered to remove genes that are not significantly

10   detectable in any of the samples as determined to be Present by the Absolute Call decision matrix. The data were further filtered to remove genes that had not been Increased or Decreased by any of the treatments as determined by the Difference Call decision matrix of GeneChip® Analysis Suite Version 3.0. After these two filtering processes, 1434 genes were left. These genes were subject to additional reduction of dimensions.

15   It is found that standard clustering software (such as that found in thestatistical package S-Plus from Mathsoft, Inc. [www.mathsoft.com], which was used for most of this analysis in this example) was still swamped by noise in the large number of genes. Further reduction in dimension was required. A variety of techniques were tried to isolate a smaller number of genes that might show differences in behavior between the agonist and antagonist

20   treatments. In some instance, genes in which the variable called Average Difference Intensity of Experiments was highest among the various treatments (*i.e.*, the genes with the highest mean A.D.I.E. or highest expression) were selected for cluster analysis. Genes for which the variable Fold Change had highest variance across the different treatments were also selected for cluster analysis. With either of these, good clustering (using S-Plus's hclust, a standard

25   hierarchical clustering routine) was performed.

          Genes whose Fold Change variable behaved most differently between the agonist and antagonist treatments were also examined. The distance used was the Kolmogorov-Smirnov distance between two empirical distributions, which works as follows. At any threshold value for fold change, it could be determined what fraction of the Fold

30   Change values were below it and above it for either the agonist or antagonist populations. The K.-S. distance is the largest difference (across all possible thresholds) between the fractions for the agonists and antagonists. This measure provided slightly more consistent differentiation among the populations. The Kolmogorov-Smirnov distance is described in the

reference article by Wilcox (1998). "Kolmogorov-Smirnov Test". Encyclopedia of

Biostatistics (Armitage, P., and Colton, T., eds.), vol.3, pp. 2174-2176. John Wiley & Sons,

New York, NY, incorporated by reference for all purposes.

5   For visualization purposes, the first two principal components were computed

and graphed. Patterns in the data (*See*, Figs. 7 and 8), including the separation that permitted

clustering for some cases, and to see the collapse of data that made separation impossible for

other subsets.

C.   Significance of the Result.

An important aspect of the characterization of drug candidates is the prediction of

10   efficacious and toxic side effects. It is desirable to use expression profiling to identify

patterns or fingerprints that signify specific drug effects, including side effects that can

indicate toxicity. Specifically, this example shows that, using the method of the invention

and software implementation of the method, known drug effects and properties, such as

activation mode and target specificity, can be correlated with expression patterns that are

15   likely to be the signature for a specific drug effect or property. One of skill in the art would

appreciate that this approach may be employed to identify drug toxicity resulted from

undesirable side effects of the drugs.

The analysis showed that agonists and antagonists can be clustered into separate

groups based on the behavior of the most variable genes, *i.e.,* activation mode (agonists

20   versus antagonists) and target specificity ($\alpha 1$, $\alpha 2$, $\beta 1$, and $\beta 2$ adrenergic receptors). Thirty-

five genes have been identified whose expression changes correlated with the activation

properties of the compounds targeted at adrenergic receptors (*See*, Figs. 7 and 8). This study

suggest that that genes with expression patterns correlated to specific compounds properties

can be potentially used as fingerprints for the drug properties, *e.g.*, activation mode, target

25   specificity, efficacy, toxicity, structure, etc.

Conclusion

The present inventions provide greatly improved methods for analyzing gene

expression profiles. It is to be understood that the above description is intended to be

30   illustrative and not restrictive. Many variations of the invention will be apparent to those of

skill in the art upon reviewing the above description. By way of example, the invention has

been described primarily with reference to the use of a high density oligonucleotide array, but

it will be readily recognized by those of skill in the art that other nucleic acid arrays, other

methods of measuring transcript levels and gene expression monitoring at the protein level could be used. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

5

10

We claim:

1.   A method for analyzing a plurality of biological profiles comprising:

a.   providing said plurality of biological profiles, wherein each of said profiles comprises a plurality of biological variables;

5   b.   reducing the number of said biological variables to obtain a set of reduced profiles; and

c.   subjecting said reduced profiles to statistical analysis.

2.   The method of Claim 1 wherein said plurality of biological variables comprises more
10   than 100 variables.

3.   The method of Claim 2 wherein said plurality of biological variables comprises more than 1000 variables.

15 4.   The method of Claim 3 wherein said plurality of biological variables comprises more than 2000 variables.

5.   The method of Claim 1 wherein said reducing step comprises selecting variables from said biological variables based upon the degree of variation of said biological
20   variables among said profiles.

6.   The method of Claim 5 wherein said degree of variation is the variance of said biological variables among said profiles.

25 7.   The method of Claim 5 wherein said degree of variation is the fold change of said biological variables among said profiles.

8.   The method of Claim 1 wherein said reducing step comprises selecting variables from said biological variables based upon the level of said biological variables among said
30   profiles.

9.   The method of Claim 1 wherein said reducing step comprises selecting variables from said biological variables based upon the degree of variation of said biological

variables among said profiles.

10.   The method of Claim 9 wherein said degree of variation is the variance of said biological variables among said profiles.

5

11.   The method of Claim 9 wherein said degree of variation is the fold change of said biological variables among said profiles.

12.   The method of Claim 1 wherein said statistical analysis is cluster analysis.

10

13.   The method of Claim 12 wherein said statistical analysis is a hierarchical cluster analysis.

14.   The method of Claim 1 wherein said statistical analysis is principal component

15   analysis.

15.   The method of Claim 1 wherein said biological profiles are gene expression profiles and each of said biological variables represents the expression of a gene.

20   16.   The method of Claim 15 wherein said expression is measured using a DNA probe array.

17.   A computer program product comprising:
       a.   computer code that receives a plurality of biological profiles, each of said
25   profiles comprises a plurality of biological variables;
       b.   computer code that reduces the number of said biological variables to obtain a set of reduced profiles;
       c.   computer code that performs statistical analysis of said reduced profiles; and
       d.   a computer readable medium that stores the computer codes.

30

18.   The computer program product of Claim 17 wherein said computer code for reducing said number of said biological variables comprises computer code that selects said reduced set of variables from said biological variables based upon the degree of

variation of said biological variables among said profiles.

19. The computer program product of Claim 18 wherein said degree of variation is the variance of said biological variables among said profiles.

5

20. The computer program product of Claim 17 wherein said degree of variation is the fold change of said biological variables among said profiles.

21. The computer program product of Claim 17 wherein said computer code for reducing said number of said biological variables comprises computer code that selects the reduced set of variables from said biological variables based upon the level of said biological variables among said profiles.

22. The computer program product of Claim 21 wherein said computer code for reducing said number of said biological variables further comprising code that selects variables from said biological variables based upon the degree of variation of said biological variables among said profiles.

23. The computer program product of Claim 22 wherein said degree of variation is the variance of said biological variables among said profiles.

24. The computer program product of Claim 22 wherein said degree of variation is the fold change of said biological variables among said profiles.

25. The computer program product of Claim 17 wherein said computer code for statistical analysis comprises computer code that performs cluster analysis.

26. The computer program product of Claim 25 wherein said cluster analysis is hierarchical cluster analysis.

27. The method of Claim 17 wherein said computer code for statistical analysis comprises computer code that performs principal component analysis.

28.   A method of studying a plurality of drugs comprising:

     a.   measuring the expression of more than 50 genes in a biological sample in response to said plurality of drugs to obtain a plurality of gene expression profiles, each of said profiles representing the response of said biological sample to one of said plurality of drugs;

     b.   reducing said profiles by selecting a plurality of genes from said at least 50 genes to obtain a set of reduced gene expression profiles; and

     c.   performing a statistical analysis using said reduced gene expression profiles.

29.   The method of Claim 28, wherein said measuring step measures the expression of at least 1000 genes.

30.   The method of Claim 28, wherein said measuring step measures the expression of at least 2000 genes.

31.   The method of Claim 28, wherein said measuring step measures the expression of at least 4000 genes.

32.   The method of Claim 28, wherein statistical analysis is a cluster analysis.

33.   The method of Claim 28, wherein said cluster analysis is a hierarchical cluster analysis.

34.   The method of Claim 28 wherein said reducing step comprises selecting genes from said at least 50 genes based upon the degree of variation of expression of said genes among said profiles.

35.   The method of Claim 34 wherein said degree of variation is the variance of expression of said genes among said profiles.

36.   The method of Claim 34 wherein said degree of variation is the fold change of expression of said at genes among said profiles.

37.   The method of Claim 28 wherein said reducing step comprises selecting genes based upon the level of expression of said genes among said profiles.

38.   The method of Claim 37 wherein said reducing step further comprises selecting genes from said genes based upon the degree of variation of expression of said genes among said profiles.

39.   The method of Claim 38 wherein said degree of variation is the variance of the expression of said genes among said profiles.

40.   The method of Claim 38 wherein said degree of variation is the fold change of the expression of said genes among said profiles.

41.   A method for classifying a plurality of drugs comprises the method of Claims 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, wherein said cluster analysis is performed using said genes as variables (Y axis).

42.   A method for classifying a plurality of genes comprises the method of Claims 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, wherein said cluster analysis is performed using said drugs as variables (Y axis).

43.   A method for classifying a plurality of drugs comprises Claims 28, 29, 30, 31, 34, 35, 36, 37, 38, 39, or 40, wherein said statistical analysis is a principal component analysis.

44.   The method of Claim 43 wherein said method further comprising a step of display said selected genes in a surface comprising a first axis representing the first component of said principal analysis and a second axis representing the second component of said principal analysis, wherein said first axis is perpendicular to said second axis.

45.   A computer program product comprising:

a.   computer code that receives a plurality of biological profiles, each of said profiles comprises a plurality of biological variables;

e.   computer code that reduces the number of said biological variables to obtain a set of reduced profiles;

f.   computer code that performs a principal component analysis of said reduced profiles;

g.   computer code that displays said biological variables according a first axis and a second axis, wherein said first axis is the first component of said principal component analysis and said second axis is the second component of said principal component analysis, and wherein said first axis is perpendicular to said second axis; and

h.   computer readable medium that stores the computer codes.

46.   The computer program product of Claim 45 wherein said computer code for reducing said number of said biological variables comprises computer code that selects said reduced set of variables from said biological variables based upon the degree of variation of said biological variables among said profiles.

47.   The computer program product of Claim 46 wherein said degree of variation is the variance of said biological variables among said profiles.

48.   The computer program product of Claim 46 wherein said degree of variation is the fold change of said biological variables among said profiles.

49.   The computer program product of Claim 45 wherein said computer code for reducing said number of said biological variables comprises computer code that selects the reduced set of variables from said biological variables based upon the level of said biological variables among said profiles.

50.   The computer program product of Claim 49 wherein said computer code for reducing said number of said biological variables further comprising code that selects variables from said biological variables based upon the degree of variation of said biological

variables among said profiles.

51.   The computer program product of Claim 50 wherein said degree of variation is the variance of said biological variables among said profiles.

5

52.   The computer program product of Claim 50 wherein said degree of variation is the fold change of said biological variables among said profiles.

10
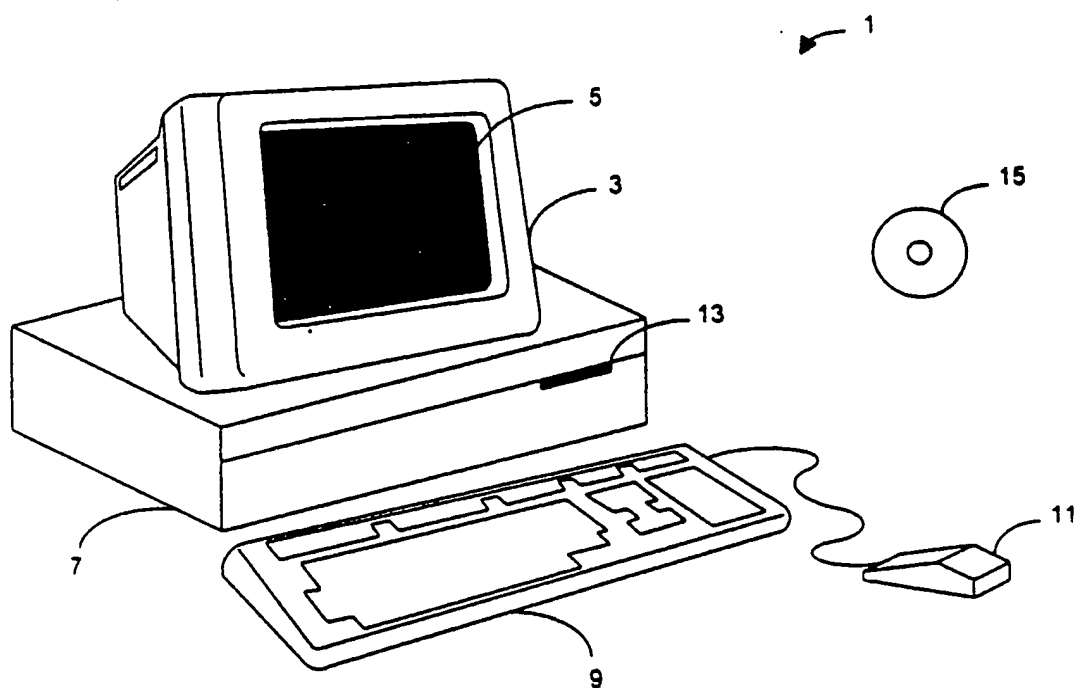
FIG. 1
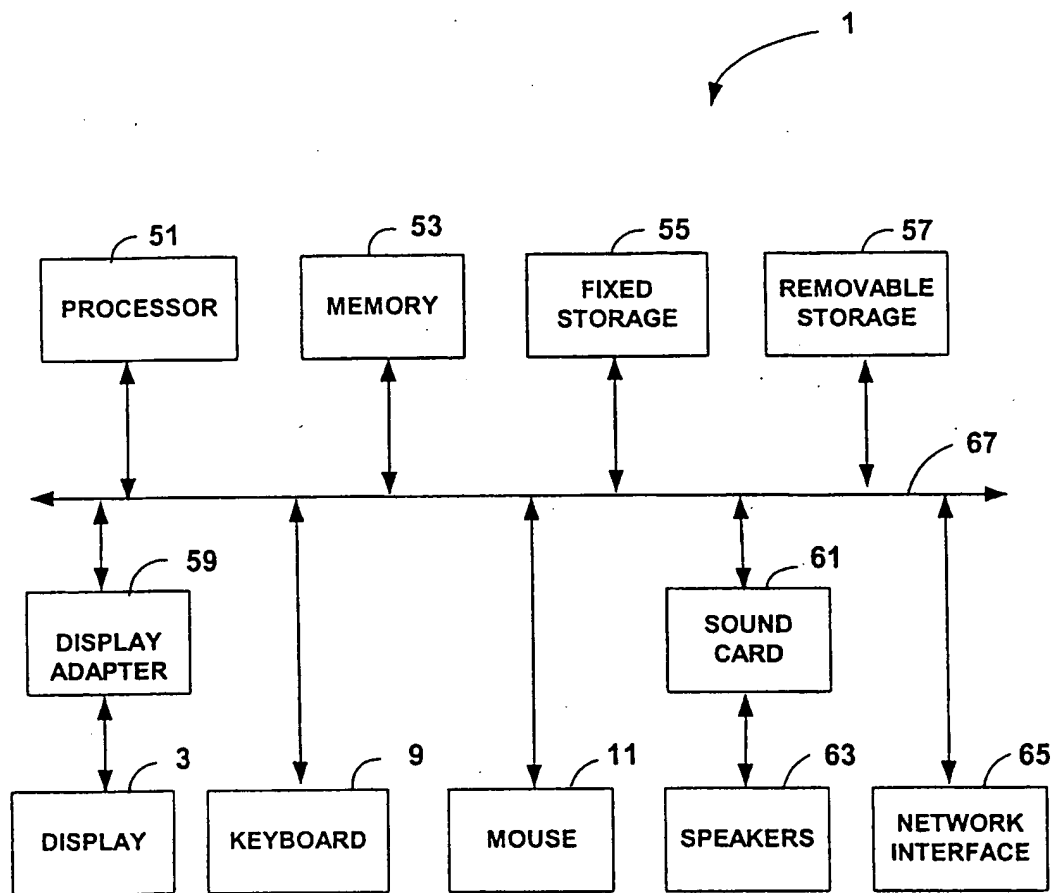
FIG. 2
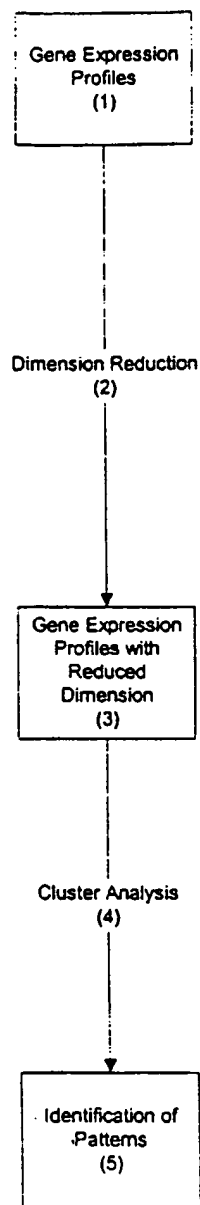
Figure 3. A Process for Gene Expression
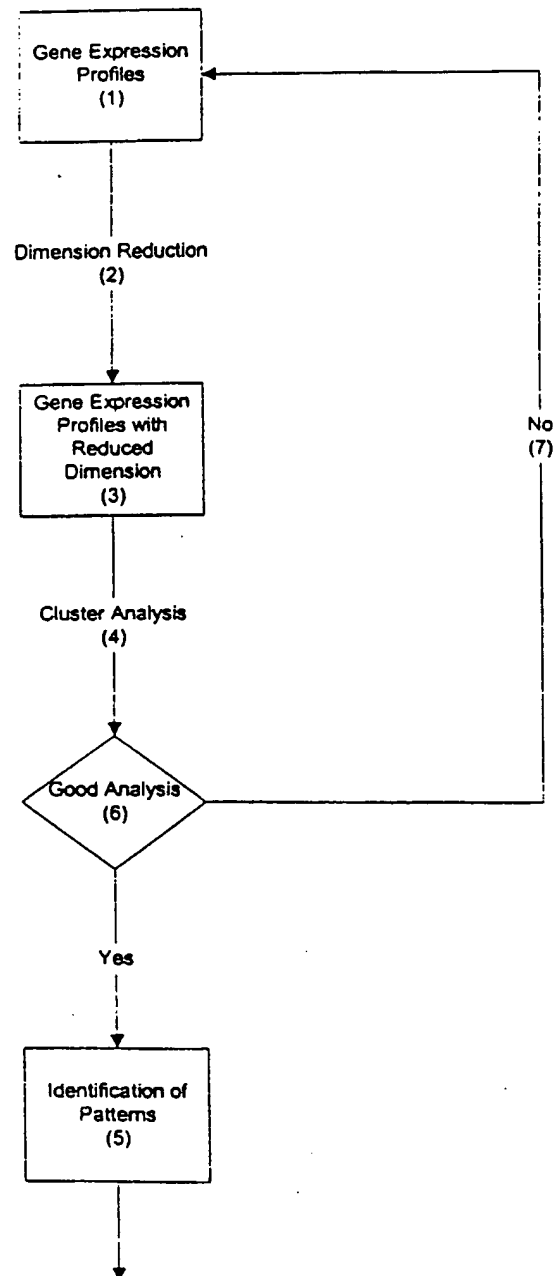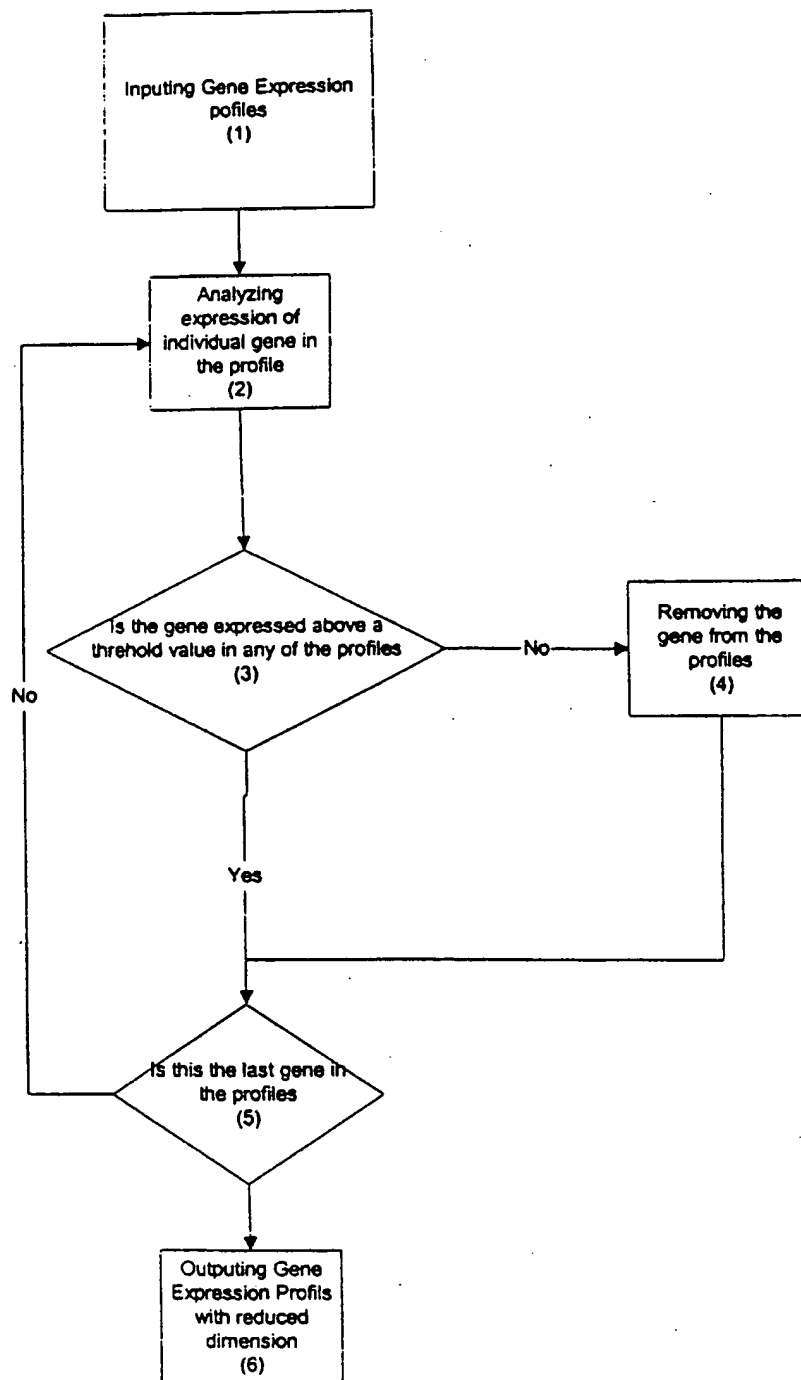Profile Analysis

Figure 4. An Iterative Process for Gene
Expression Profile Analysis

Figure 5. A Process for Reduction of Dimension of Gene
Expression Profiles

Figure 6

Figure 7

Crosscheck Compound Clusters

Most of the agonists can be separated from antagonists in this data set by Principal components analysis based on behavior of the 35 genes

Figure 8

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :C12Q 1/68
US CL :435/6

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6,69.1; 514/2,44; 536/22.1,23.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS, MEDLINE, WPI, BIOTECH ABS, EMBASE covering search terms: profiles, expression, statistics, correlate, and gene

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 5,830,645 A (PINKEL et al.) 03 November 1998, see especially the abstract.	1 & 5-12 ------------ 2-4 and 13-52
X --- Y	US 5,800,992 A (FODOR et al.) 01 September 1998, see the entire document.	1 & 5-12 ------------ 2-4 and 13-52
X --- Y	US 5,635,400 A (BRENNER) 03 June 1997, see the entire document.	1 & 5-12 ------------ 2-4 and 13-52

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 JANUARY 2001	21
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer DELLA MAE COLLINS ARDIN MARSCHEL PARALEGAL SPECIALIST Telephone No. (703) TECHNOLOGY CENTER 1600

Form PCT/ISA/210 (second sheet) (July 1998)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 5,654,413 A (BRENNER) 05 August 1997, see entire document.	1 & 5-12 ---------- 2-4 and 13-52
X --- Y	US 5,338,659 A (KAUVAR et al.) 16 August 1994, see especially the abstract.	1 & 5-12 ---------- 2-4 and 13-52
X --- Y	US 5,674,688 A (KAUVAR et al.) 07 October 1997, see entire document.	1 & 5-12 ---------- 13-52
X --- Y	ALTSCHUL, S.F., "A Protein Alignment Scoring System Sensitive at All Evolutionary Distances", Journal of Evolution, 1993, Volume 36, pages 290-300, see entire document.	1 & 5-12 ---------- 2-4 and 13-52
X --- Y	"Description of Piroutte Algorithms", Chemometrics Techical Note, Published by Infometrix, Ind., P.O. Box 1528, Woodinville, WA, issued 1993, pages 1-4, see entire document.	1 & 5-12 ---------- 2-4 and 13-52